题目：非洲猪瘟病毒基因组中DNA互补回文模式的分布特征研究

作者：许肖枫[①]，陈泽[①]，罗建勋[①]，刘光远[①]，任巧云[①]，罗金[①]，殷宏[①*]，高山[②*]

① 中国农业科学院兰州兽医研究所家畜疫病病原生物学国家重点实验室，甘肃省动物寄生虫病重点实验室，江苏省动物重要疫病与人兽共患病防控协同创新中心，兰州 730046;

② 南开大学生命科学学院，天津 300071;

* 联系人, E-mail: gao_shan@mail.nankai.edu.cn; yinhong@caas.cn

资助项目：中央级公益性科研院所基本科研业务费

摘要 受互补回文小RNA发现的启发，本研究从小RNA水平考虑DNA互补回文模式的生物学功能，并揭示了非洲猪瘟病毒基因组中较长（14 bp以上）DNA互补回文模式的分布特征．本研究中公开的非洲猪瘟病毒基因组中的DNA互补回文模式可以用于设计引物或探针，提高病毒检测的灵敏度和特异度；也可以直接用于设计小干扰RNA（small interfering RNA，siRNA）进行RNA干扰实验．这样，不需要病毒感染，更多的基础研究工作者可以使用这种间接方法研究非洲猪瘟病毒感染和致病机制．本研究提供的研究病毒基因组中DNA互补回文模式的思路和方法，可以推广到其他微生物或动植物基因功能或进化研究领域，具有重要的理论意义．

关键词：非洲猪瘟病毒 回文小RNA 互补回文小RNA DNA互补回文模式 小干扰RNA **（通常5~8个）**

回文序列（palindromic sequence），也叫DNA回文模式（palindromic DNA motif）是一段倒置重复序列，其广泛存在于各种生物体基因组中．生物功能已知的DNA回文模式包括限制性酶切位点、甲基化位点和T细胞受体相关序列等 [1]。经典的DNA回文模式要求DNA两条链从5’到3’方向的序列完全一致，不同于经典定义，本研究重新定义了DNA回文模式和DNA互补回文模式．受限制性酶切位点模式的启发，当前对DNA互补回文模式的研究主要集中于较短（一般不超过10 bp）的模式；对于DNA回文模式的研究鲜有报道；而对于动植物基因组中大量存在的大片段DNA回文模式和DNA互补回文模式，其生物学功能依然处于未知状态．2018年，南开大学高山和中国农业科学院兰州兽医研究所陈泽等在国际上首次报道了SARS病毒（severe acute respiratory syndrome coronavirus，SARS-CoV）中存在互补回文小RNA（complemented palindromic small RNA, cpsRNA）的现象；并通过进化与分子功能分析相结合的方法，初步证明了互补回文小RNA可能在SARS病毒感染或致病方面起作用 [1]．该研究首次从小RNA水平考虑DNA互补回文模式的生物学功能，并揭示了病毒基因组中较长（14 bp以上）DNA互补回文模式的重要特征（**见结果**）．这些特征的分析为研究病毒感染和致病机制提供了新的思路和方法．非洲猪瘟病毒(African Swine Fever Virus, ASFV [2])于2018年在中国首次发现，由于它能导致一种急性发热、传染性高且致死率高达100%的猪瘟 [3]，其感染和致病机制的研究亟待开展。本研究应用我们前期工作 [1]中提出的思路和方法，对非洲猪瘟病毒基因组中DNA互补回文模式的特征进行了统计分析．并发现非洲猪瘟病毒基因组中DNA互补回文模式的基因组密度明显高于其他双链DNA病毒，一些DNA互补回文模式与此前发现的SARS病毒基因组中得到功

能实验验证的DNA互补回文模式高度相似. 这些发现为后续的非洲猪瘟病毒感染和致病机制研究提供了新的思路和方法.

1　方法与数据

本研究中所用到的 8 种病毒（HPV-18、HBV、HCV、HIV-1、EBV、SMRV、SARS-CoV 和 ASFV）参考基因组序列（KU298886.1、JQ688404.1、D11168.1、KM390026.1、M80517.1、M23385.1、DQ497008.1 和 FN557520.1）来自 NCBI GenBank 数据库. 人类基因 GAPDH 序列（ENSG00000111640.14）来自 Ensembl 基因组数据库. 非洲猪瘟病毒小 RNA 高通量测序数据(SRA: ERP018944)来自 NCBI SRA 数据库. 高通量测序数据质量控制等使用软件 Fastq_clean v2.0[4]; 测序数据比对到病毒参考基因组使用软件 Bowtie v0.12.7; 统计与作图使用软件 R v2.15.3[5]; 比对结果校对等使用软件 Tablet v1.15.09.01[6]. DNA 互补回文模式的二级结构的最小自由能（Minimum Free Energy, MFE）的计算使用在线服务 RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi), DNA 互补回文模式的茎环结构的链内退火温度 T_m 的计算使用公式为 4(G+C)+2(A+T). 统计病毒基因组中 DNA 回文模式或 DNA 互补回文模式数量时，要求茎部长度不小于 7 bp，环部长度为 0-5 bp. DNA 互补回文模式的基因组密度的计算公式为某个基因组中全部 DNA 互补回文模式的数量乘以 1000 后再除以基因组长度。

2　结果

我们重新定义了DNA回文模式和DNA互补回文模式, 主要基于两点: 1. 前期研究暗示了两者具有不同的生物学意义; 2. 在动物病毒基因组中, 两者分布具有不同的特征. DNA回文模式要求DNA两条链从5’到3’方向的序列互补配对（例如ATCGGCTA）；DNA互补回文模式与经典的DNA回文模式定义相同（例如ATCGCGAT）. 2017年，南开大学卜文俊和高山等在国际上首次报道了人类线粒体基因组D-loop区转录的两条长非编码RNA(long non-coding RNA, lncRNA)的全长序列和首个发现的回文小RNA(palindromic small RNA, psRNA)，并推测小RNA长度的DNA回文模式可能与转录调控有关 [7]. 2015年，南开大学高山等在使用小RNA高通量测序数据研究病毒特征序列过程中，无意中发现了一系列互补回文小RNA来自SARS病毒，这些序列是DNA互补回文模式TCTTTAACAAGCTTGTTAAAGA（**见图1A**）产生的；后期对互补回文小RNA SARS-CoV-cpsR-19（**见图1A**）的RNA干扰实验结果显示该小RNA可以引起显著性的细胞凋亡. 由于自身具备的特殊序列模式，回文小RNA与互补回文小RNA被定义为一类新小RNA，但是两者没有生物学意义上的关联. 由于DNA互补回文模式可能与病毒感染或致病有关，本研究仅考虑病毒基因组中的DNA互补回文模式，对于DNA回文模式不做更多讨论。

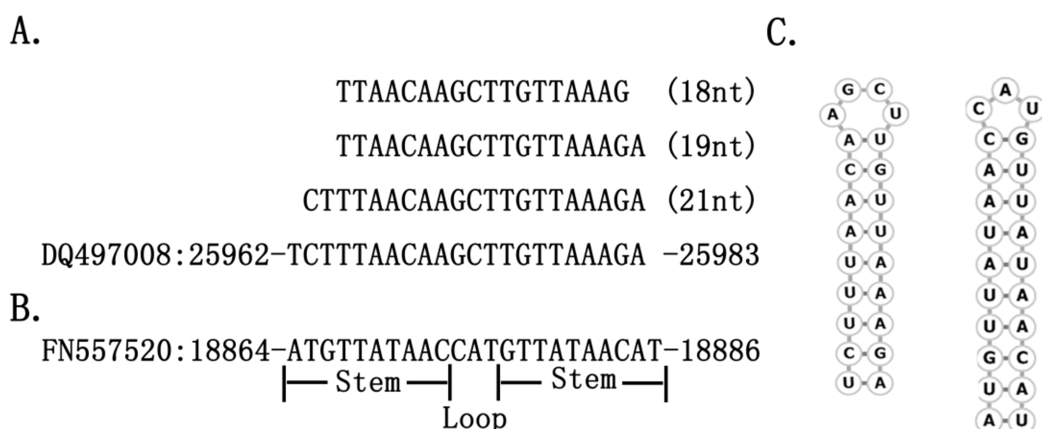


图1 DNA互补回文模式与互补回文小RNA

Figure 1 Complemented palindromic DNA motif and complemented palindromic small RNA

A. SARS病毒(GenBank: DQ497008.1)中发现的互补回文小RNA，长度分别为18, 19（命名为SARS-CoV-cpsR-19）和21 nt; **B.** DNA互补回文模式的重新定义，此序列来自表1第2条; **C.** SARS病毒(GenBank: DQ497008.1)中DNA互补回文模式（左）与非洲猪瘟病毒(GenBank: FN557520.1)中DNA互补回文模式（右），此序列来自表1第1条和第2条. 用DNA互补回文模式计算得到的对应RNA的二级结构叫做该DNA互补回文模式的二级结构，这个二级结构不是这段DNA序列真实存在的二级结构，其计算得到的最小自由能等属性属于该DNA对应RNA的二级结构，因此序列中的T可以用U表示。

当前对DNA互补回文模式的研究主要集中于较短（一般不超过10 bp）的模式，而且只考虑严格定义（即模式中所有碱基都参与互补配对）的情况，因此存在很大的局限性。 Chew等在国际上首次报道了SARS病毒基因组中DNA互补回文模式的规律 [8]，其主要发现包括两点：1. 在所有分析的冠状病毒基因组中，长度为4 bp的DNA互补回文模式出现频率显著较低；2. 长度为6 bp的DNA互补回文模式仅在SARS病毒（而不是所有冠状病毒）基因组中频率显著较低. 因此，该研究的结论是SARS病毒基因组含有较少的6 bp DNA互补回文模式有可能利于该病毒有效规避宿主细胞内的某些防御机制而有利于病毒存活；该研究同时还发现了两个较长（14 bp以上）的DNA互补回文模式分别是TCTTTAACAAGCTTGTTAAAGA和TAAAATTAATTTTA。根据两个概率模型的计算结果，这两个DNA互补回文模式的出现不是偶然的. 根据经典定义，Chew等仅仅在SARS病毒基因组中找到了两个较长的DNA互补回文模式；受互补回文小RNA发现的启发，我们将DNA互补回文模式的定义放宽要求，并在SARS病毒基因组中多找到了27个较长的DNA互补回文模式 [7]。

参考RNA发卡结构（hairpin）的定义，将DNA回文模式和DNA互补回文模式从序列上分为组成回文或互补回文的茎（stem）和不参与组成的环（loop）两部分（**见图1B**）。我们同时定义，用DNA互补回文模式计算得到的对应RNA的二级结构（**见方法与数据**）叫做该DNA互补回文模式的二级结构，这个二级结构不是这段DNA序列真实存在的二级结构，其计算得到的最小自由能等属性属于该DNA对应RNA的二级结构。DNA回文模式和DNA互补回文模式中的茎环结构的定义是基于DNA序列的，与二级结构无关，DNA互补回文模式茎部的全部碱基在计算得到的二级结构中不一定参与碱基互补配对。DNA互补回文模式的二级结构的最小自由能的计算与对应RNA

发卡结构的最小自由能的计算相同，都要考虑非经典配对GU（图形显示时可以表示为GU或GT）. DNA互补回文模式的茎环结构的链内退火温度的计算不考虑其对应的二级结构，而是使用公式4(G+C)+2(A+T)计算一侧茎部内全部碱基（**见表1**）.

放宽要求后，DNA互补回文模式的环部可以允许至多5个碱基存在（此参数仅适用于病毒基因组）。经过对8种哺乳动物病毒基因组的统计分析，我们发现哺乳动物病毒基因组中较长（14 bp以上）的DNA互补回文模式具有三个非常典型的特征：1. DNA互补回文模式数量随长度显著递减，超过一定长度（31 bp）的模式不出现，即截断效应；2. 仅有极少量DNA互补回文模式可以从当前数据库中搜索到对应的互补回文小RNA，即DNA互补回文模式产生互补回文小RNA的概率很低（可能由于技术原因漏检）. 3. 部分DNA互补回文模式在进化上高度保守，仅有少量转换发生，而且这些突变都不影响其二级结构的稳定性. 为了比较不同病毒基因组中DNA互补回文模式的含量，我们定义了DNA互补回文模式的基因组密度（**见方法与数据**）。我们发现不同病毒基因组中DNA互补回文模式的基因组密度差异显著，同一种病毒不同亚型或株系之间的差异也很明显。HPV-18（双链DNA病毒）、HBV（双链DNA病毒）、HCV（正链RNA病毒）、HIV-1（正链RNA病毒）、EBV（双链DNA病毒）、SMRV（单链RNA病毒）、SARS-CoV（正链RNA病毒）和ASFV（双链DNA病毒）中DNA互补回文模式的基因组密度分别是 0.64(1000*5/7857)、0.93(1000*3/3215)、0.95(1000*9/9436)、0.21(1000*2/9709)、0.62(1000*115/184113)、0.68(1000*6/8785)、0.98(1000*29/29727)和1.1(1000*199/181187). 在正链RNA病毒中，HIV-1基因组中DNA互补回文模式的基因组密度明显低于其他几种病毒，而且还明显低于人类基因GAPDH中DNA互补回文模式的基因组密度0.45 (1000*2/4448)；而在双链DNA病毒中，DNA互补回文模式的基因组密度从HPV-18到HBV再到ASFV差异很大。非洲猪瘟病毒（ASFV）基因组中DNA互补回文模式的基因组密度高达1.1，其DNA互补回文模式数量达到199个，其中有16个（**见表1**）与此前发现的SARS病毒基因组中的DNA互补回文模式高度相似（**见图1C**）.

表1 非洲猪瘟病毒基因组中16个DNA互补回文模式

Figure 1 Selected complemented palindromic DNA motifs in the ASFV genome

Complemented palindromic DNA motif	Start	End	Length	Loop	GC %	Tm
TCTTTAACAAGCTTGTTAAAGA*	25962	25983	22	0	27.27	20
ATGTTATAACCATGTTATAACAT	18864	18886	23	3	21.74	24
TTATGACAAAACATGTCATAA	26019	26039	21	5	23.81	20
TTGTATACAAAGGTATACAA	43837	43856	20	4	25	20
TATCACAATTGCGATACAATTGTGATA	47875	47901	27	5	29.63	28
ACAATTGTGATACAATTGT	47890	47908	19	3	26.32	20
ACAATTGTGATACAATTGT	47901	47919	19	3	26.32	20
AAAACTTTTCGAGAAAAGTTTT	57175	57196	22	2	22.73	24
CATATCTAATAGTAGATATG	59625	59644	20	4	25	20
TTGCAAACAAATATTTGTTTGCAA	65604	65627	24	0	25	30
TATTACGGTCTTTTACCGTAATA	77602	77624	23	5	30.43	24
TAAACGTTTAAACTAAACGTTTA	81734	81756	23	3	21.74	24

GTTTAAACTAAACGTTTAAAC	81739	81759	21	5	23.81	20
CTCTTTTTTGGAAAAAAAAGAG	140565	140586	22	4	27.27	22
TTTGAAATCAGCGATTTCAAA	141328	141348	21	3	28.57	22
TTTTCCAAAATGTTTGGAAAA	172636	172656	21	3	23.81	22
TGTTGTTACAAACAACA	178692	178708	17	3	29.41	18

*表示的DNA互补回文模式来自SARS病毒基因组(GenBank: DQ497008.1)，其编码的首个互补回文小RNA SARS-CoV-cpsR-19能引起显著的细胞凋亡[1]. 猪瘟病毒基因组(GenBank: FN557520.1)中16个DNA互补回文模式与SARS-CoV-cpsR-19各种属性相似, T_m表示DNA互补回文模式的茎环结构的链内退火温度, MFE表示DNA互补回文模式的二级结构的最小自由能.

对于病毒基因组中DNA互补回文模式的统计分析，我们重点考察长度、GC含量、链内退火温度和最小自由能四个属性的分布特征. 三种病毒（EBV、SARS-CoV和ASFV）基因组中DNA互补回文模式的长度分布总体上满足长度显著递减，超过一定长度（31 bp）的模式不出现（**见图1A**）. SARS-CoV在17 bp互补回文模式上出现最大值; EBV在19 bp和23 bp互补回文模式上出现两个反常的升高, 根据这两个反常我们发现了EBV编码的一个微小RNA（microRNA, miRNA）[9]. 三种病毒中DNA互补回文模式的GC含量分布代表了三个类型（**见图1B**）, 分别是偏高（EBV），中等（SARS-CoV）和偏低（ASFV）。三种病毒中DNA互补回文模式的链内退火温度分布总体上集中在18和20度（**见图1BC**），ASFV在14和16度有反常，EBV在24和30度有反常. 20度正好是小RNA建库的温度，在这个温度建库，接头很可能因为互补回文小RNA的二级结构而没有连接成功，因此导致大量的互补回文小RNA在小RNA高通量测序中被漏检。三种病毒中DNA互补回文模式的最小自由能分布代表了三个类型，分别是偏高（ASFV），中等（SARS-CoV）和偏低（EBV）。这些分布特征反映了病毒的一些内在特性，对于病毒序列分析具有非常重要的意义.

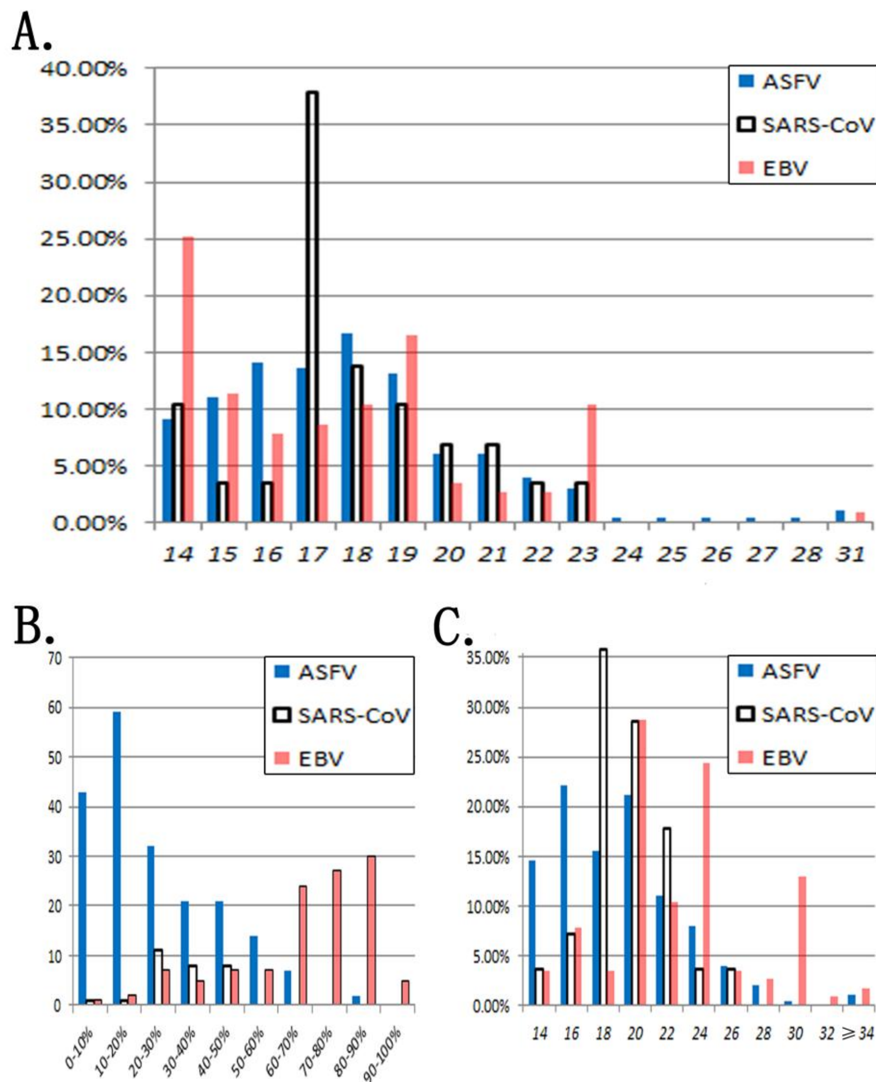


图2 DNA互补回文模式的分布特征

Figure 2 Distributive characteristics of complemented palindromic DNA motifs

A. 三种病毒基因组中DNA互补回文模式的长度分布. **B.** 三种病毒基因组中DNA互补回文模式的GC含量分布. **C.** 三种病毒基因组中DNA互补回文模式的链内退火温度分布.

3 讨论和结论

本研究从小RNA水平考虑DNA互补回文模式的生物学功能，并揭示了非洲猪瘟病毒基因组中较长（14 bp以上）DNA互补回文模式的重要特征. 尽管从当前的NCBI SRA数据库中仅有的一组小RNA高通量测序数据(SRA: ERP018944）中没有搜索到来自非洲猪瘟病毒基因组中的DNA互补回文模式区域的互补回文小RNA，但是我们的分析结果依然支持这些模式与非洲猪瘟病毒感染和致病机制相关。本研究的意义包括：提供了一种方法研究病毒基因组中的DNA互补回文模式; 可以根据DNA互补回文模式的保守性设计引物或探针，提高病毒检测的灵敏度和

特异度；特别是，本研究公开的DNA互补回文模式可以直接用于设计小干扰RNA（small interfering RNA, siRNA）进行RNA干扰实验. 这样，不需要病毒感染，更多的基础研究工作者可以使用这种间接方法研究非洲猪瘟病毒感染和致病机制.

作为双链DNA病毒的非洲猪瘟病毒和作为正链RNA病毒的SARS病毒都具有典型的DNA互补回文模式的统计性特征，但是非洲猪瘟病毒能否产生互补回文小RNA，这些小RNA的产生机制以及是否具有生物学功能等这些问题还有待继续研究。根据当前研究经验，作为正链RNA病毒的SARS病毒很可能在形成RNA双链过程中被宿主体内防御机制（例如RNA）切割而产生互补回文小RNA，作为双链DNA病毒的非洲猪瘟病毒转录的RNA也可以在内部形成双链部分，但是能否在单链内切割产生互补回文小RNA依然未知. 互补回文小RNA产生后，由于其独特的茎环结构，是否能够继续被切割而产生更小（7~13）片段的小RNA以及是否具有生物学功能也是一个重要的研究方向。我们推测，很可能存在这些更小的小RNA参与反转录的引发或宿主基因的转录调控.

致谢：感谢南开大学生命科学学院卜文俊教授、南开大学数学科学学院阮吉寿教授和南开大学医学院刘畅老师对本研究工作的长期支持，感谢南开大学生命科学学院硕士研究生牛晓冉、姬海硕和金秀峰参与部分工作。

参考文献

1 Liu C, Ze Chen, Hu Y, Ji HS, Yu DS, Shen WY, Li SY, Ruan JS, Bu WJ; Gao, S (2018) Complemented palindromic small RNAs first discovered from SARS Coronavirus. Genes : in press

2 Montgomery RE (1921) On a form of swine fever occurring in British East Africa (Kenya Colony). Journal of comparative pathology and therapeutics 34:159-191

3 Galindo I, Alonso C (2017) African swine fever virus: a review. Viruses 9(5):103

4 Zhang M, Sun H, Fei Z, Zhan F, Gong X, Gao S Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality control. In: Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on, 2014. IEEE, pp 44-48

5 Gao S, Ou J, Xiao K (2014) R language and Bioconductor in bioinformatics applications(Chinese Edition). Tianjin Science and Technology Translation and Publishing Co., Tianjin

6 Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D (2012) Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform:bbs012.

7 Shan G, Tian X, Sun Y, Wu Z, Cheng Z, Dong P, Zhao Q, He B, Ruan J, Bu W (2017) Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data. Mitochondrion 38:41-47.

8 Chew, DSH; Choi, KP; Heidner, H; Leung, MY (2004) Palindromes in SARS and other coronaviruses. Informs J. Comput. 16:331 - 340.

9 Wang F; Sun Y; Ruan JS; Chen R; Chen X; Chen CJ; Kreuze JF; Fei ZJ; Zhu X; Gao S (2016) Using small RNA deep sequencing to detect human viruses. BioMed Res. Int. 2016, 2016, 2596782.